DOI:10.11931/guihaia.gxzw201907038

基于高通量测序的都匀地区福鼎大白种茶树根茎叶分析

王芬*，裴会敏，文狄，陈志，刘荣，姚玉仙，马媛
（黔南民族师范学院 生物科学与农学院 贵州 都匀 558000）

摘要：茶树富含儿茶素、茶氨酸和咖啡碱等多种具有广泛保健功能的植物代谢产物，为探究茶树中茶多酚等产物代谢途径的相关基因，本研究以贵州都匀地区福鼎大白种茶树的根、茎、叶为研究对象，利用高通量测序技术构建茶的转录组数据库并筛选其根、茎、叶差异表达基因。共获得 70.88 Gb Clean Data，各样品 Clean Data 均达到 6.33 Gb，Q30 碱基百分比在 93.22%以上。将 Clean Reads 与中国种茶树参考基因组进行序列比对，比对效率从 87.83%到 91.14%。基于比对结果，进行可变剪接预测分析和基因结构优化分析，发掘新基因 13 531 个，其中 10 244 个得到功能注释。利用 FPKM 进行基因表达量分析，根据基因在不同样品中表达量识别差异表达基因。叶与茎的差异基因有 5 595 个，其中 2 769 个在茎中上调，2 826 下调，叶和根有 9 650 个差异基因，5 056 个上调，4 594 个下调，茎和根中有 5 644 个差异基因，2 938 个上调，2 706 个下调，并通过 GO 和 KEGG 分析，将差异基因进行功能注释和富集分析。研究结果为揭示都匀地区福鼎大白种茶参与类黄酮、茶氨酸和咖啡碱等代谢途径相关的基因提供参考，为选育优良品种等提供理论依据。

关键词：都匀地区，福鼎大白种，根茎叶，高通量测序，差异基因
中图分类号：Q945　　　　　　　　　　　　**文献标识码**：A

High-throughput sequencing analysis of root, stem and leaf in Fudingdabai

基金项目：国家自然科学基金（31900486）；贵州省科技厅基础研究计划项目（黔科合基础[2019]1298）；贵州省教育厅青年科技人才成长项目（黔教合 KY 字[2018]420）；黔南民族师范学院科研创新团队项目（Qnsyk201605, QNYSKYTD2018011）；黔南民族师范学院高层次人才引进研究专项项目（qnsyrc201611）；黔南州科技计划项目（黔南科合学科建设农学（2018）5 号和黔南科合农字（2017）23）；贵州省教育厅（黔教合人才团队字[2015]68）；黔南民族师范学院重大科研创新基金博士专项基金（QNSY2018BS018）；贵州省植物学重点支持学科开放基金（qnsyzw1809）；黔南民族师范学院继续教育基地项目（QNSY2018ZJ006）[Supported by the National Natural Science Foundation of China(31900486); Guizhou Provincial Science and Technology Foundation((QKH Basic research) [2019]1298); Guizhou Provincial Education Department(QJH-KY-Z [2018]420); Research and Innovation Team of Qiannan Normal University for Nationalities(Qnsyk201605, QNYSKYTD2018011); Scientific Research Project of Qiannan Normal University for Nationalities(qnsyrc201611); Qiannan Science and Technology Bureau(QNKHXKJSNX(2018)5 and [QNKHNZ(2017)23]; Guizhou Provincial Education Department([2015]68); Scientific Research Project of Qiannan Normal University for Nationalities(QNSY2018BS018); Scientific Research Project of Qiannan Normal University for Nationalities(qnsyzw1809); Scientific Research Project of Qiannan Normal University for Nationalities (QNSY2018ZJ006)]。

作者简介：王芬（1984-），女，山西运城人，博士，讲师，主要从事茶树转录组学研究，(E-mail) fenmin521@163.com。
*通信作者

WANG Fen*, PEI Huimin, WEN Di, CHEN Zhi, LIU Rong, YAO Yuxian, Ma Yuan

(*The Department of Life Science and Agriculture, Qiannan Normal University for Nationalities*, Duyun 558000, Guizhou, China)

Abstract: Tea tree is rich in catechins, theanine, caffeine and other metobolite of health fuction. In order to study the related genes of the metabolisms of the polyphenols. We use high-throughput sequencing technology to study the root, stem and leaf of Fudingdabai tea and find differential expression genes. The results showed that 70.88 Gb Clean Data was obtained, 6.33 Gb Clean Data is in each sample and Q30 is more than 93.22%. We map the Clean Reads to reference genome, the blast result is from 87.83% to 91.14%. Then, alternative splicing and gene structure optimization was analyzed. There are 13 531 new genes, in which, 10 244 genes were annotated. GO and KEGG functional annotation and enrichment analysis were carried out in differential expression genes, which were identified according to gene expression level in different samples. There were 5 595 DEGs between leaf and stem, 2 769 genes were up-regulated and 2 826 genes were down-regulated. 9 650 DEGs were found beween leaf and root, 5 056 genes were up-regulated and 4 594 genes were down-regulated. 5 644 genes between stem and root, 2 938 genes were up-regulated and 2 706 genes were down-regulated. The results are expected to provide reference for recognizing genes of catechins, theanine, caffeine pathways, provide the theoretical basis for breeding improved seeds.

Key words: Duyun, Fudingdabai, root-stem-leaf, high-throughput sequencing, differential expression gene

都匀地处云贵高原苗岭山脉南侧，具有低纬度、高海拔、寡日照、多云雾的自然生态环境，适宜茶树生长，资源丰富（陈世军等，2017）。不同品种的适制性、抗逆性和生长势不同，将导致茶品质和产量的不同。90 年代以来，都匀市等地开始大量引种单产高、品质好、适应性强的福鼎大白种茶树，增加了茶树的遗传多样性。目前，关于都匀地区的茶种质资源保护和抗性研究薄弱（张丽娟等，2016），特别是在转录组层面上的研究较少，在一定程度上限制了茶产业的应用发展。而福鼎大白种茶园占全市茶园面积 85%以上，因此，开展都匀地区福鼎大白种根、茎、叶的转录组研究对茶树在品质、质量及抗逆相关的遗传改良上具有重要意义。

随着高通量测序技术的出现，由于该技术能快速全面获取研究对象在某一状态下基因转录信息，因此该技术广泛应用于生物体转录组基因表达分析，能够准确发掘重要功能基因。Shi et al.（2011）用 RNA-seq 技术对龙井 43 的嫩叶、成熟叶、茎、幼根、花蕾以及成熟种子进行测序，总共获得 127 094 个基因，其中包括与茶特有香气相关的代谢产物如咖啡碱、茶氨酸、黄酮等合成相关的基因。王君雅等（2019）利用转录组技术对龙井 43 和中茶 126 进行差异基因表达分析，结果表明 99 个表达模式完全相反的基因参与的生物过程主要有 MAPK 信号通路、谷胱甘肽和苯丙烷等。利用转录组测序技术可识别编码次生代谢的调节基因和结构基因，也可用来预测未知基因的功能，是进行基因组功能研究的重要工具（Novaes et al., 2008）。本研究利用高通量测序技术，对都匀地区福鼎大白种茶树根、茎、叶中类黄酮相关代谢途径进行研究，探索根、茎、叶中差异表达基因参与的代谢途径，为探讨都匀地区福鼎大白种茶树生长发育及组织间功能差异的分子机制提供理论依据。

1 材料与方法

1.1 材料

供试材料选用黔南州都匀地区种植的茶树-福鼎大白种。茶苗选用黔南民族师范学院茶园的扦插苗。选取 9 株培养条件相同长势相同的茶苗分为 3 组，每组 3 株，在第 1 组中取嫩根、嫩茎和嫩叶分别放入 3 个离心管作为第 1 组生物学重复的样本，分别标为根的第 1 个生物学重复，茎的第 1 个生物学重复，叶的第 1 个生物学重复。每个样本 3 个生物学重复，剩余的两组生物学重复分别在剩下的两组茶苗中取样，共 9 个样本，分别为叶 1、叶 2、叶 3、茎 1、茎 2、茎 3、根 1、根 2 和根 3。采样时在茶树上剪取发育阶段相似、部位相同、充分伸展的叶片、茎和根，样品随后迅速放入液氮罐中速冻后放入-80 ℃冰箱储存备用。

1.2 转录组测序

将采集的福鼎大白种茶苗根茎叶样品放在干冰中送往北京百迈客生物科技有限公司进行二代 Illumina 高通量测序，对测序数据进行质量控制，并将二代得到的转录本与已测序的中国种茶树基因组通过序列比对的方法进行比较。然后进行转录组文库质量评估、SNP/InDel 分析、可变剪接事件预测、差异基因分析和差异外显子分析等。

Illumina 二代测序数据提交至 NCBI 的 SRA 数据库，BioProject 的编号为 PRJNA562747。

2 结果与分析

2.1 测序数据质量分析

从茶苗上分别取福鼎大白的叶、茎和根作为实验材料进行转录组测序，碱基质量值达到 Q30 的占 93.22%以上，GC 含量为 44.56%~45.85%（表 1），结果表明转录组文库质量符合分析要求。

表 1 样品数据产出统计表

Table1 Output statistics among samples

样品名称 Samples ID	高质量 Reads 数目 Clean reads	数据量 Clean bases	GC 含量 GC content	准确率>0.999 Q30 (%)
叶 1 Leaf 1	21 828 512	6 502 479 526	45.36%	93.62%
叶 2 Leaf 2	28 450 839	8 467 453 936	45.21%	94.22%
叶 3 Leaf 3	29 057 531	8 663 984 698	45.18%	93.80%
茎 1 Stem 1	21 212 767	6 326 706 172	44.71%	93.70%
茎 2 Stem 2	25 760 743	7 695 169 918	45.35%	93.65%
茎 3 Stem 3	26 080 601	7 745 810 876	44.56%	93.79%
根 1 Root 1	23 789 995	7 078 295 558	45.85%	93.22%
根 2 Root 2	29 354 206	8 728 835 412	45.61%	93.48%
根 3 Root 3	32 496 205	9 667 694 674	45.72%	93.54%

2.2 转录组与中国种茶树基因组比对

2.2.1 比对效率

本研究利用 HISAT2（Kim et al., 2015）对测序数据与中国种茶树基因组（Wei et al., 2018）进行比对，利用 StringTie（Pertea et al., 2015）对比对上的 reads 进行组装和定量，比对到参考基因组上的 Reads 占 Clean Reads 的百分比从 87.83-91.14%（表 2）。

表 2 转录组与基因组比对

Table 2 Transcriptome blast to genome

样品 Samples	Clean Reads 数 Total Reads	比对到参考基因组上的 Reads 数及在 Clean Reads 中占的百分比	比对到参考基因组唯一位置的 Reads 数目及在 Clean Reads 中占的百分	比对到参考基因组多处的 Reads 数及在 Cleans Reads 中占的百分比 Multiple Map Reads	比对到参考基因组正链的 Reads 数及在 Clean Reads 中占的百分比	比对到参考基因组负链的 Reads 数及在 Clean Reads 中占的百分比

		Mapped Reads	比 Uniq Mapped Reads		Reads Map to '+'	Reads Map to '-'
叶 1 Leaf 1	43 657 024	39 632 673 (90.78%)	37 386 481 (85.64%)	2 246 192 (5.15%)	19 481 885 (44.62%)	19 587 096 (44.87%)
叶 2 Leaf 2	56 901 678	51 861 422 (90.09%)	48 949 427 (86.02%)	2911995 (5.12%)	25 493 617 (44.80%)	25 604 077 (45.00%)
叶 3 Leaf 3	58 115 062	52 353 177 (90.09%)	49 391 581 (84.99%)	2 961 596 (5.10%)	25 733 339 (44.28%)	25 849 934 (44.48%)
茎 1 Stem 1	42 425 534	38 353 796 (90.40%)	36 275 113 (85.50%)	2 078 683 (4.90%)	18 802 471 (44.32%)	18 912 508 (44.58%)
茎 2 Stem 2	51 521 486	46 959 156 (91.14%)	44 300 646 (85.98%)	2 658 510 (5.16%)	22 986 304 (44.61%)	23 138 767 (44.91%)
茎 3 Stem 3	52 161 202	46 917 710 (89.95%)	44 231 706 (84.80%)	2 686 004 (5.15%)	22 960 018 (44.02%)	23 121 829 (44.33%)
根 1 Root 1	47 579 990	41 789 100 (87.83%)	38 194 971 (80.28%)	3 594 129 (7.55%)	19 697 394 (41.40%)	20 253 048 (42.57%)
根 2 Root 2	58 708 412	52 846 174 (90.01%)	49 318 371 (84.01%)	3 527 803 (6.01%)	25 605 318 (43.61%)	25 872 980 (44.07%)
根 3 Root 3	64 992 410	58 353 149 (89.78%)	54 585 798 (83.99%)	3 767 351 (5.80%)	28 312 409 (43.56%)	28598 694 (44.00%)

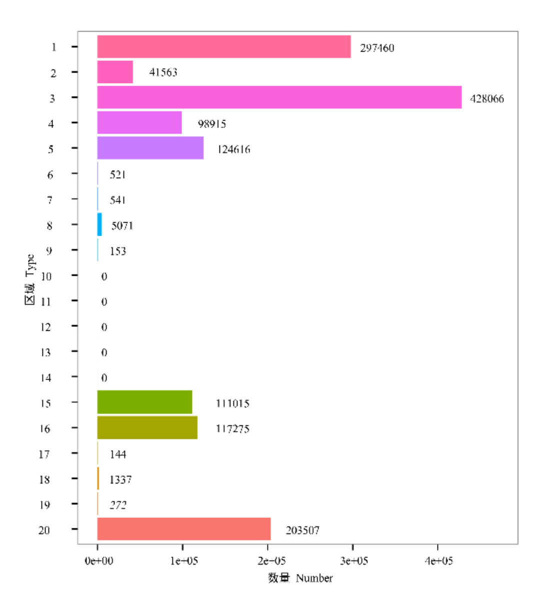
2.2.2 SNP/InDel 分析

利用 GATK（Mckenna et al., 2010）软件对 Hisat2 比对结果中的 SNP 位点和 InDel 进行识别，进而分析基因表达水平和基因功能，SNP 分析表明基因区 SNP 位点数多于基因间区，转换型 SNP 多于颠换型 SNP（表 3）。InDel 主要存在于内含子区和基因间区（图 1）。

表 3 SNP 位点统计表

Table 3 SNP loci statistics

样品 Sample	SNP 数目 SNP Number	基因区 SNP 位点数 Genic SNP	基因间区 SNP 位点数 Intergenic SNP	转换型 SNP Transition SNP(%)	颠换型 SNP Transversion SNP(%)	杂合型 SNP Heterozygosity(%)
叶 1 Leaf 1	540 707	428 327	112 380	66.07%	33.93%	51.38%
叶 2 Leaf 2	541 852	413 786	128 066	66.25%	33.75%	34.86%
叶 3 Leaf 3	583 002	448 343	134 659	66.33%	33.67%	45.95%
茎 1 Stem 1	489 642	393 947	95 695	65.46%	34.54%	49.63%
茎 2 Stem 2	504 670	394 472	110 198	65.89%	34.11%	35.85%
茎 3 Stem 3	638 706	496 194	142 512	65.90%	34.10%	50.79%
根 1 Root 1	521 166	409 410	111 756	65.74%	34.26%	50.91%
根 2 Root 2	523 515	396 706	126 809	66.22%	33.78%	34.70%
根 3 Root 3	582 747	447 260	135 487	66.14%	33.86%	48.58%



1. 基因间；2. 基因内；3. 内含子；4. 上游；5. 下游；6. 受体剪切位点；7. 供体剪切位点；8. 剪切位点；9. 起点缺失；10. 移码；11. 密码子缺失；12. 密码子插入；13. 密码子改变和缺失；14. 密码子改变和插入；15. 同义编码区；16. 非同义编码区；17. 同义终止区；18. 终止区；19. 终止缺失；20. 其他。

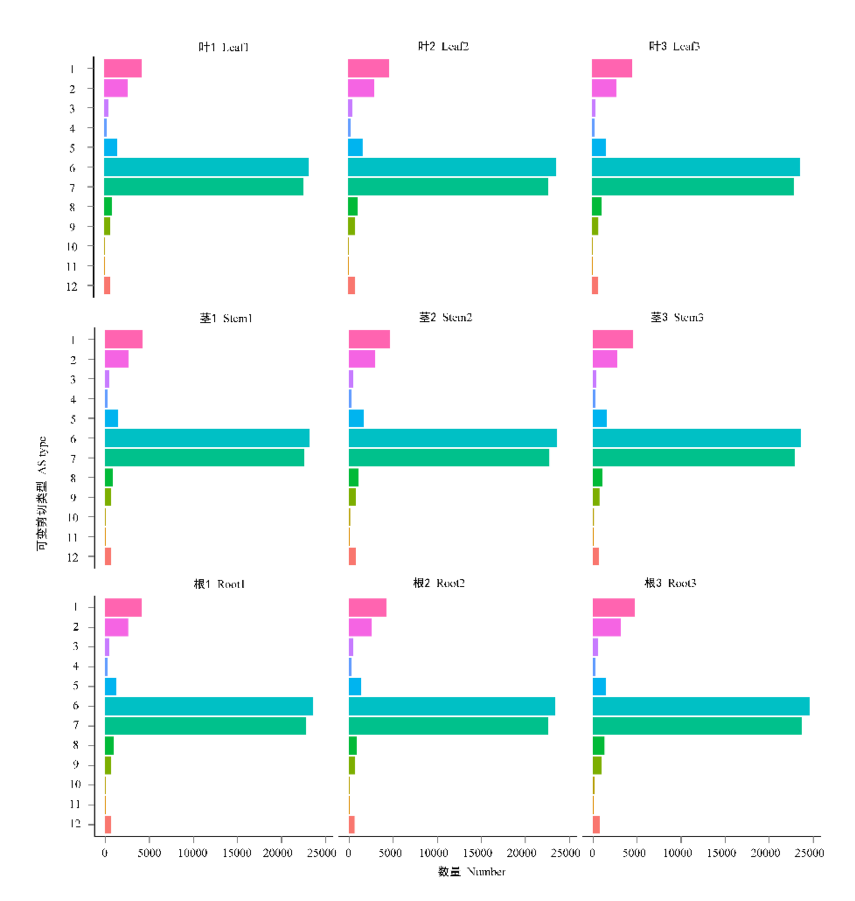
1. Intergenic; 2. Intragenic; 3. Intron; 4. Upstream; 5. Downstream; 6. Splice site acceptor; 7. Splice site donor; 8. Splice site region; 9. Start lost; 10. Frame shift; 11. Condon deletion; 12. Condon insertion; 13. Condon change plus condon deletion; 14. Condon change plus condon insertion; 15. Synonymous coding; 16. Non synonymous; 17. Synonymous stop; 18. Stop gained; 19. Stop lost; 20. Other.

图 1 InDel 注释分类

Fig. 1 InDel annotation

2.2.3 可变剪接预测

基因通过转录生成前体 mRNA，再经过不同的剪接，产生不同的成熟 mRNA，翻译为不同的蛋白质。利用 Asprofil（Florea et al., 2013）软件对可变剪接类型和表达量进行分析，结果表明在 9 个样品中 5’端和 3’端外显子可变剪切最多（图 2）。



1. 可变 5'或 3'端剪切；2. 单内含子滞留；3. 多内含子滞留；4. 多外显子跳跃；5. 单外显子跳跃；6. 第一个外显子可变剪切；7. 最后一个外显子可变剪切；8. 可变 5'或 3'端剪切（模糊边界）；9. 单内含子滞留（模糊边界）；10. 多内含子滞留（模糊边界）；11. 多外显子跳跃（模糊边界）；12. 单外显子跳跃（模糊边界）。

1. Ae; 2. Ir; 3. Mir; 4. Mskip; 5. Skip; 6. Tss; 7. Tts; 8. Xae; 9. Xir; 10. Xmir; 11. Xmskip; 12. Xskip.

图 2 可变剪接事件数量统计

Fig.2 Alternative splicing

2.2.4 基因结构优化

由于处理数据的软件不同或数据本身的局限性，该文对中国种茶树基因组的 261 个基因结构进行了优化并列出了部分优化的基因（表 4）。

表 4 优化的基因

Table 4 Gene structure optimized

基因 ID Gene ID	基因座 Locus	正负链 Strand	优化位置 Site	原来注释的第一个或最后一个外显子起止坐标 Original first or termination region	优化后第一个或最后一个外显子起止坐标 Optimized first or termination region
TEA004484.1_gene	Scaffold1014:163897-231030	+	5'	205 536-205 536	163 897-205 536
TEA000973.1_gene	Scaffold1019:1350887-1371675	+	3'	1 371 277-1 371 277	1 371 277-1 371 675
TEA016708.1_gene	Scaffold1045:1434954-1455906	+	5'	1 435 240-1 435 240	1 434 954-1 435 240
TEA004051.1_gene	Scaffold1054:303905-308542	+	3'	308 318-308 318	308 318-308 542
TEA000171.1_gene	Scaffold1058:112241-129629	-	5'	129 534-129 534	129 534-129 629
TEA007140.1_gene	Scaffold1067:122675-156025	-	3'	123 878-123 878	122 675-123 878
TEA020092.1_gene	Scaffold1093:635489-640729	+	3'	640 607-640 607	640 607-640 729
TEA020094.1_gene	Scaffold1093:532337-549457	-	5'	549 397-549 397	549 397-549 457
TEA020098.1_gene	Scaffold1093:1162317-1191258	-	3'	1 162 515-1 162 515	1 162 317-1 162 515

TEA001055.1_gene	Scaffold1099:848040-870252	-	3'	848 142-848 142	848 040-848 142

2.3 新基因

2.3.1 新基因功能注释

我们对利用 StringTie 拼接出来的转录本与中国种茶树基因组的注释信息进行比较，发掘出了 13 531 个新基因，并将它们与 Swiss-Prot, COG, Pfam, KEGG, GO 和 NR 数据库进行 BLAST，获得新基因的注释信息（表 5）。

表 5 新基因功能注释结果

Table 5 Annotation of new genes

注释数据库 Annotated databases	新基因数目 New gene number
COG	2 363
GO	6 218
KEGG	3 948
KOG	5 833
Pfam	5 383
Swiss-Prot	6 900
eggNOG	9 243
nr	10 107
All	10 244

2.3.2 基因表达量

该文通过最大流量算法，采用 FPKM（Fragments Per Kilobase of transcript per Million fragments mapped）作为衡量基因表达量高低的标准，FPKM 计算公式如下：

$$\mathrm{FPKM} = \frac{\mathrm{cDNA\ Fragments}}{\mathrm{Mapped\ Fragments(Millions)} \times \mathrm{Transcript\ Length(kb)}}$$

福鼎大白茶根茎叶三个部位的基因表达量的密度分布对比（图 3）表明，能够测序到的编码蛋白的基因 FPKM 值跨越 10-^{2} 到 10^{4} 六个数量级。

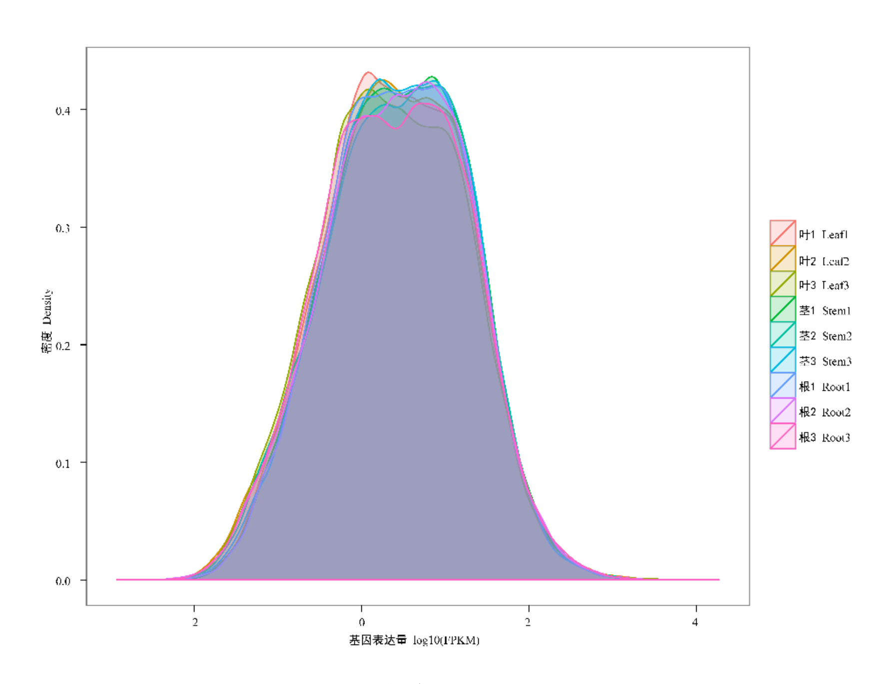


图 3 FPKM 密度分布对比

Fig.3 FPKM density

2.4 差异表达分析

2.4.1 差异表达筛选

利用 Deseq（Wang et al., 2010）进行样品间的差异表达分析，并且将 Fold Change≥2，FDR<0.01 作为筛选标准，该文列出了部分的差异表达基因并且统计了根茎叶两两样品间的差异基因的数目（表 6，表 7）。从叶和茎，叶和根，茎和根的差异表达火山图（图 4）中可以看出差异基因的统计学显著性以及上调基因和下调基因，红点代表上调，绿点代表下调。此外，我们还计算了三个组织差异表达基因的聚类热图（图 5）。

表 6 差异表达基因

Table 6 Differential expression genes

基因 ID GeneID	错误发现率 FDR	表达量差异倍数对数值 log2FC	上调或下调 Regulated
TEA000001.1_gene	0.000944095	2.088239721	Up
TEA000014.1_gene	1.00E-11	-2.156608862	Down
TEA000018.1_gene	3.88E-39	-4.204470344	Down
TEA000020.1_gene	4.20E-20	-2.852144505	Down
TEA000025.1_gene	0.000388739	2.729470094	Up
TEA000028.1_gene	1.06E-10	7.71789907	Up
TEA000040.1_gene	3.80E-08	4.115902534	Up
TEA000053.1_gene	1.02E-12	-8.291825123	Down
TEA000057.1_gene	2.37E-12	1.911663895	Up

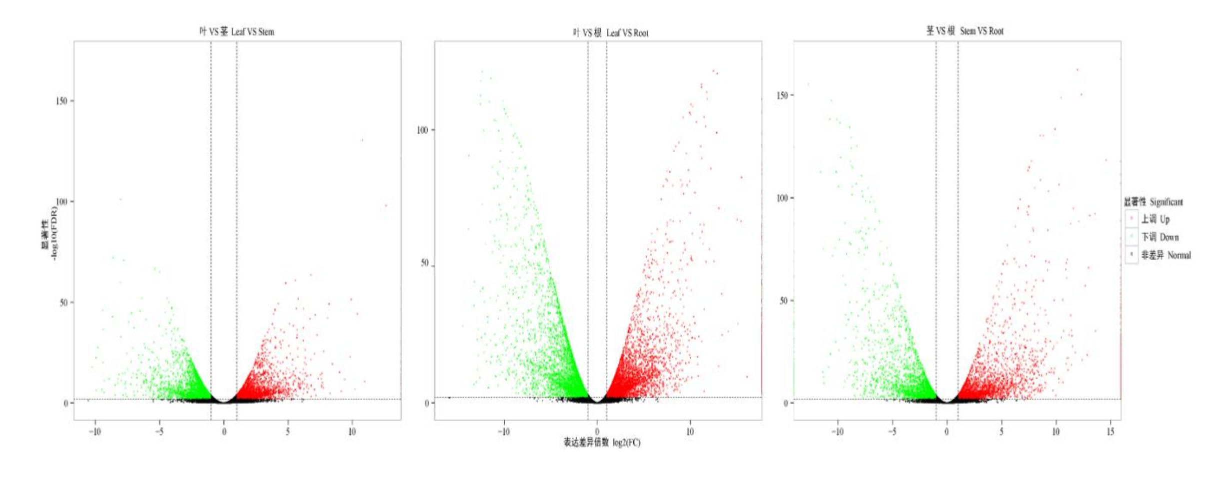


图 4 差异表达火山图

Fig. 4 Volcano plot

表 7 差异表达基因数目

Table 7 DEGs number

差异基因集 DEG Set	差异基因数目 DEG Number	上调基因数目 Up-regulated	下调基因数目 Down-regulated
叶 VS 茎 Leaf VS Stem	5 595	2 769	2 826
叶 VS 根 Leaf VS Root	9 650	5 056	4 594
茎 VS 根 Stem VS Root	5 644	2 938	2 706

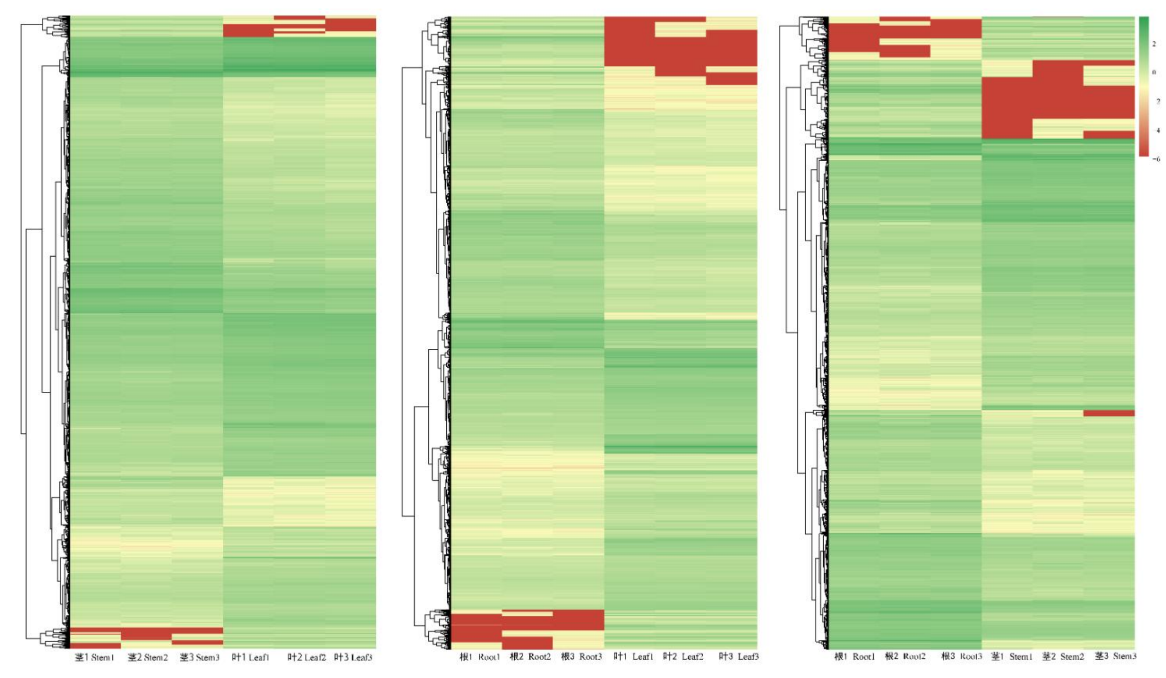


图 5 差异表达基因聚类

Fig.5 Heatmap of DEGs

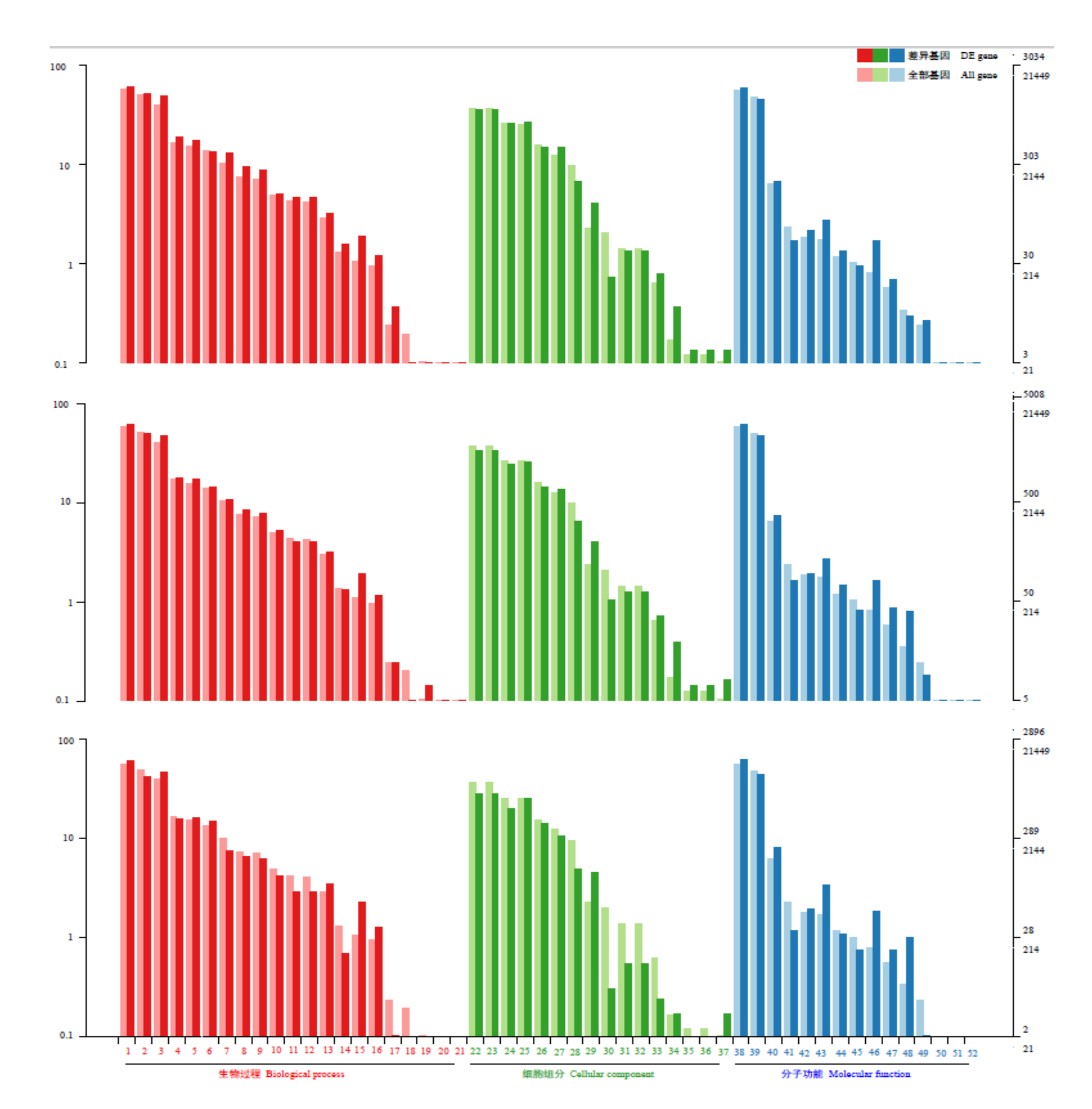
2.4.2 差异基因功能注释和富集分析

对差异基因进行 COG, GO, KEGG, KOG, NR, Pfam, Swiss-Prot 和 eggNOG 功能注释（表 8），并且绘制了叶和茎、叶和根、茎和根差异基因的 GO 功能分类图（图 6）。应用超几何检验，对差异表达基因进行 KEGG 通路富集分析，统计了显著性 Q 值最小的前 20 个通路（图 7），图中每个圆圈代表一个 pathway 通路，圆圈大小代表通路总富集的基因数目，圆圈越大，表示基因越多。圆圈颜色代表 q value，q value 越小，代表差异基因在该通路中的富集性越可靠，所以说该图越靠近右下角的的圆圈，参考价值越大。

表 8 注释的差异表达基因数量统计表

Table 8 Annotation of DEGs in database

差异基因集 DEG Set	注释到的所有差异基因数目 Total	注释到 COG 的差异基因 COG	注释到 GO 的差异基因 GO	注释到 KEGG 的差异基因 KEGG	注释到 KOG 的差异基因 KOG	注释到 NR 的差异基因 NR	注释到 Pfam 的差异基因 Pfam	注释到 Swissprot 的差异基因 Swiss-Prot	注释到 eggNOG 的差异基因 eggNOG
叶 VS 茎 Leaf VS Stem	5 444	2 476	3 019	1 972	2 738	5 381	4 662	4 520	5 337
叶 VS 根 Leaf VS Root	9 331	4 255	4 966	3 357	4 897	9 199	7 900	7 654	9 083
茎 VS 根 Stem VS Root	5 454	2 633	2 876	1 899	2 768	5 363	4 780	4 629	5 298



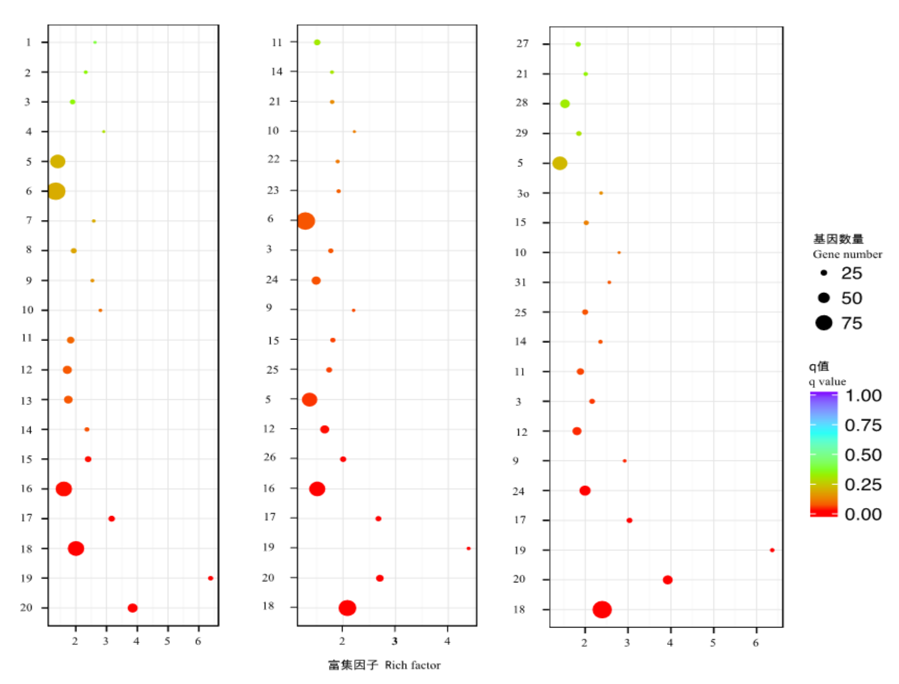
横坐标注：1. 代谢过程；2. 细胞过程；3. 单组织过程；4. 生物调节；5. 应激反应；6. 定位；7. 细胞成分组织或生物合成；8. 发育过程；9. 多细胞生物过程；10. 信号；11. 生殖；12. 生殖过程；13. 多组织过程；14. 生长；15. 解毒作用；16. 免疫系统过程；17. 生物附着；18. 生物相；19. 节律性过程；20. 细胞杀伤性；21. 运动；22. 细胞；23. 细胞组件；24. 细胞器；25. 细胞膜；26. 细胞膜组件；27. 细胞器组件；28. 大分子复合物；29. 胞外区；30. 细胞膜内控；31. 细胞连接；32. 共质体；33. 超分子复合物；34. 拟核；35. 病毒粒子；36. 病毒粒子组件；37. 胞外区组件；38. 催化活性；39. 结合；40. 转运活性；41. 结构分子活性；42. 核酸结合转录因子活性；43. 电子载体活性；44. 信号转导活性；45. 分子功能调节因子；46. 抗氧化活性；47. 分子转导活性；48. 营养库活性；49. 转录因子活性，蛋白结合；50. 蛋白标签；51. 金属伴侣活性；52. 翻译调控活性。

Abscissa note: 1. Metabolic process; 2. Cellular process; 3. Single-organism process; 4. Biological regulation; 5. Response to stimulus; 6. Localization; 7. Cellular component organization or biogenesis; 8. Developmental process; 9. Multicellular organismal process; 10. Signaling; 11. Reproduction; 12. Reproductive process; 13. Multi-organism process; 14. Growth; 15. Detoxification; 16. Immune system process; 17. Biological adhesion; 18. Biological phase; 19. Rhythmic process; 20. Cell killing; 21. Locomotion; 22. Cell; 23. Cell part; 24. Organelle; 25. Membrane; 26. Membrane part; 27. Organelle part; 28. Macromolecular; 29. Extracellular region; 30. Membrane-enclosed lumen; 31. Cell junction; 32. Symplast; 33. Supramolecular complex; 34. Nucleoid; 35. Virion; 36. Virion part; 37. Extracellular region part; 38. Catalytic activity; 39. Binding; 40. Transporter activity; 41. Structural molecule activity; 42. Nucleic acid binding transcription factor activity; 43. Electron carrier activity; 44. Signal transducer activity; 45. Molecular function regulator; 46. Antioxidant activity; 47.

Molecular transducer activity; 48. Nutrient reservoir activity; 49. Transcription factor activity， protein binding; 50. Protein tag; 51. Metallochaperone activity; 52. Translation regulator activity.

图 6 差异表达基因 GO 注释分类

Fig.6 GO functional classification



纵坐标注：1. 有机含硒化合物代谢；2. 类胡萝卜素生物合成；3. 类黄酮生物合成；4. 糖尿病并发症中的年龄信号通路；5. 淀粉和蔗糖代谢；6. 碳代谢；7. 叶酸碳池；8. 半乳糖代谢；9. 角质、亚氨酸和蜡的生物合成；10. 单萜类生物合成；11. 甘氨酸、丝氨酸和苏氨酸代谢；12. 光合生物固碳作用；13. 乙醛酸和二羧酸代谢；14. 氮代谢；15. 氰基氨基酸代谢；16. 植物激素信号转导；17. 卟啉和叶绿素代谢；18. 苯丙素的生物合成；19. 光合作用-天线蛋白；20. 光合作用；21. 抗坏血酸和醛酸代谢；22. 硫代谢；23. 双醌和其它萜烯醌生物合成；24. 谷胱甘肽代谢；25. 萜类骨架生物合成；26. 磷酸戊糖途径；27. 脂肪酸降解；28. 糖酵解/糖新生；29. 亚麻酸代谢；30. 二萜生物合成；31. 倍半萜和三萜生物合成。

Ordinate note: 1. Selenocompound metabolism; 2. Carotenoid biosynthesis; 3. Flavonoid biosynthesis; 4. AGE-RAGE signalling pathway in diabetic complications; 5. Starch and sucrose metabolism; 6. Carbon metabolism; 7. One carbon pool by folate; 8. Galactose metabolism; 9. Cutin, suberine and wax biosynthesis; 10. Monoterpenoid biosynthesis; 11. Glycine, serine and threonine metabolism; 12. Carbon fixation in photosynthetic organism; 13. Glyoxylate and dicarboxylate metabolism; 14. Nitrogen metabolism; 15. Cyanoamino acid metabolism; 16. Plant hormone signal transduction; 17. Porphyrin and chlorophyll metabolism; 18. Phenylpropanoid biosynthesis; 19. Photosynthesis-antenna proteins; 20. Photosynthesis; 21. Ascorbate and aldarate metabolism; 22. Sulfur metabolism; 23. Biquinone and other terpenoid-quinone biosynthesis; 24. Glutathione metabolism; 25. Terpenoid backbone biosynthesis; 26. Pentose phosphate pathway; 27. Fatty acid degradation; 28. Glycolysis/Gluconeogenesis; 29. Alpha-Linolenic acid metabolism; 30. Diterpenoid biosynthesis; 31. Sesquiterpenoid and triterpenoid biosynthesis.

图 7 差异表达基因 KEGG 通路富集

Fig. 7 KEGG pathway enrichment

2.5 DEU 分析

该文应用 DEXSeq（Anders et al., 2012）进行外显子水平的差异分析，即 DEU（differential exon usage）分析，设置 FDR<0.01，该文列出了叶和茎的部分 DEU（表 9）。

表 9 差异外显子表达分析

Table 9 Differential exon expression

基因 ID Gene ID	外显子 ID Exon ID	表达量差异倍数 log2(FC)	P 值 P value	错误发现率 FDR
TEA018663.1_gene	E010	0.403936012	1.04E-07	2.12E-05
TEA018663.1_gene	E011	-1.732604999	6.69E-25	9.75E-22
TEA018662.1_gene	E001	-0.96779822	5.47E-05	0.00551022
TEA012997.1_gene	E007	-0.308323557	1.95E-10	6.24E-08
TEA021927.1_gene	E001	0.168112449	1.33E-08	3.25E-06
TEA016026.1_gene	E002	-1.273705747	1.32E-05	0.001577751
TEA031666.1_gene	E001	0.032790485	1.98E-06	0.000300559
TEA031666.1_gene	E002	-0.077595321	2.13E-06	0.000320651
TEA016705.1_gene	E001	-0.600866652	1.58E-07	3.06E-05

3 讨论

福鼎大白种茶树是异花授粉的经济作物，基因组杂合度高，具有优良的发芽率、抗寒性和耐旱性（刘本英等，2008）。本研究对都匀地区福鼎大白种根茎叶进行了转录组测序，Q30 碱基百分比不小于 93.22%，样品与参考基因组的比对效率在 87.83%~91.14%之间。而李明玺等（2018）对静安白茶芽和叶的转录组研究中，Q30 值都为 87.32%，Liu et al.（2017）对瑞雪的转录组研究中，Q20 为 97%左右，Wei et al.（2018）对龙井 43 的转录组研究中，Q30 为 90%左右，因此，福鼎大白种文库构建成功且测序质量良好，可进行下一步分析。SNP 位点数目从 489 642~638 706，转换类型比例从 65.46%~66.33%,，颠换类型从 33.67%~34.54%，杂合型 SNP 位点比例从 34.70%~51.38%。对 261 个基因进行了结构优化，发现了 13 531 个新基因，其中 10 244 个新基因被注释了，注释到 COG 数据库中有 2 363 个，注释到 GO 的有 6 218 个，注释到 KEGG 的有 3 948 个，注释到 KOG 的有 5 833 个，注释到 Pfam 的有 5 383 个，注释到 SwissProt 的有 6 900 个，注释到 eggNOG 的有 9 243 个，注释到 nr 的有 10 107 个。

福鼎大白种根茎叶中有 12 595 个基因参与代谢过程， 12 507 个基因参与催化活性， 8 176 个基因参与细胞组成成分，是基因数最多的三个生物过程。陈琳波等（2015）对“紫鹃”茶树的转录组分析中，Unigene 的 GO 分析表明，代谢过程、细胞、细胞部分、催化活性等富集程度较高，与本研究结果相一致。三个组织中有 1 615 个基因参与发育过程（278，405，191），即叶和茎参与此过程的有 278 个差异基因，叶和根中有 405 个差异基因，茎和根中有 191 个。根茎叶有 1 401 个基因参与运输活动（197，356，231），288 个基因参与生长过程（47，65，20），262 个基因参与信号转导活动（44，70，30），200 个基因参与免疫反应过程（35，56，36）。

都匀地区茶独特的风味是由叶片中的茶氨酸、萜烯类和类黄酮的代谢产物赋予的。Wu et al.（2013）从茶树叶片的转录组研究中得出大多数基因参与编码合成黄酮类、咖啡碱和茶氨酸等次生代谢途径的重要酶。该文通过 KEGG 功能显著性富集分析，发现叶和茎、叶和根、茎和根差异基因参与的部分极其显著的通路有类胡萝卜素生物合成，泛醌和其他萜烯类醌的生物合成，单萜生物合成，甘氨酸、丝氨酸和苏氨酸代谢，乙醛酸和二羧酸代谢，植物激素信号转导，卟啉和叶绿素代谢，苯丙素的生物合成，光合作用天线蛋白和光合作用，类黄酮生物合成，萜类化合物的生物合成，亚麻酸代谢等显著富集。叶和茎的差异表达基因在 KEGG 通路中，有 90 个基因参与碳代谢（8.65%，

21，69），占所有差异基因的 8.65%，其中有 21 个差异基因在茎中表达上调，69 个差异基因在叶中表达上调。79 个差异基因参与植物激素信号转导（7.6%，55，24），75 个差异基因参与苯丙素的生物合成（7.21%，49，26），70 个差异基因参与淀粉和蔗糖代谢（6.73%，46，24，），64 个差异基因参与氨基酸生物合成（6.15%，22，42），43 个差异基因参与植物病原体互作通路（4.13%，30，13），19 差异基因参与类黄酮的生物合成（1.83%，13，6），20 个差异基因参与萜类化合物生物合成（1.92%，4，16），24 个差异基因参与卟啉和叶绿素代谢（2.31%，1，23），44 个差异基因参与光合作用（4.23%，0，44）。以上研究表明都匀地区福鼎大白茶独特的香味和各种健康功能主要归因于萜类物质合成、光合作用、类黄酮的生物合成等生物过程，类黄酮主要赋予茶的苦味，二萜类化合物的水解产物具有令人愉悦的香味和特色风味。为进一步研究都匀地区福鼎大白茶生长发育和遗传育种提供了分子基础，并为揭示与类黄酮相关代谢通路有关基因提供了重要理论依据。

参考文献：

ANDERS S, REYES A, HUBER W, 2012. Detecting differential usage of exons from RNA-seq data[J]. Geno Res, 22(10): 2008-2017.

CHEN LB, XIA LF, ZHOU M, et al., 2015. Mol Plant Breeding, 13(10): 2250-2255. Analysis on transcriptome sequenced for tea plant[J]. [陈琳波，夏丽飞，周萌，等，2015. 基于 RNA-Seq 技术的"紫鹃"茶树转录组分析[J]. 分子植物育种, 13(10): 2250-2255.]

CHEN SJ, ZHANGMZ, YAOYX, et al., 2017. Establishment of DNA fingerprinting for tea germplasm from qiannan prefecture by SSR markers[J]. J. Plant Gene Resour, 18(1): 106-111[陈世军，张明泽，姚玉仙,等，2017. 基于 SSR 标记的黔南茶树种质资源 DNA 指纹图谱构建[J]. 植物遗传资源学报, 18(1): 106-111.]

FLOREA L, SONG L, SALZBERG SL, 2013. Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues[J]. F1000Res, 2: 188.

KIM D, LANGMEAD B, SALZBERG SL, 2015. Hisat: A fast spliced aligner with low memory requirements[J]. Nat Methods, 12(4):357-360.

LI MX, WANG M, GAN YD, et al., 2018. Transcriptome data assembly and gene function annotation of buds and leaves of *Cammellia sinensis* cultivar Jing'anbaicha[J]. Mod Food Sci Technol, 34(5): 93-100. [李明玺，王敏，甘玉迪，等. 靖安白茶芽和叶的转录组数据组装机基因功能注释[J]. 现代食品科技, 34(5): 93-100.]

LIU BY, ZHOU J, XU M, et al., 2008. Tissue culture of immature embryo and parentage identification of hybrids between *Camellia taliensis* (w.w.Smish) melchior and *C. sinensis* 'Fuding Dabaicha'[J]. Acta Hortic Sin, 35: 735-740. [刘本英，周健，许枚，等, 2008. 云南大理茶与福鼎大白茶种间杂交幼胚的组织培养及亲子鉴定[J]. 园艺学报, 35: 735-740.]

LIU F, WANG Y, DING ZT, et al., 2017. Transcriptomic analysis of flower development in tea[J]. Gene, 631: 39-51.

MCKENNA A, HANNA M, BANKS E, et al., 2010. The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data[J]. Geno Res, 20(9): 1297-1303.

NOVAES E, DROST DR, FARMERIE WG, et al., 2008. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome[J]. BMC Geno, 9:312.

PERTEA M, PERTEA GM, ANTONESCU CM, et al., 2015. Stringtie enables improved reconstruction of a transcriptome from rna-seq reads[J]. Nat Biotechnol, 33(3): 290-295.

SHI CY, YANG H, WEI CL, et al., 2011. Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds[J]. BMC Genomics, 12:131.

WANG JY, CHEN W, LIU DD, et al., 2019. The transcriptome analysis of different tea cultivars in response to the spring cold spells[J]. J Tea Sci, 39(2): 181-192. [王君雅, 陈玮, 刘丁丁, 等, 2019. 不同品种茶树新梢响应"倒春寒"的转录组分析[J]. 茶叶科学, 39(2): 181-192.]

WANG L, FENG Z, WANG X, et al., 2010. DEGseq: An R package for identifying differentially expressed genes from RNA-seq data[J]. Bioinformatics, 26: 136-138.

WEI CL, YANG H, WANG SB, et al., 2018. Draft genome sequence of *Camellia sinensis* var. *sinensis* provides insights into the evolution of the tea genome and tea quality[J]. Proc Natl Acad Sci USA, 115(18): E4151-E4158.

WEI K, ZHANG YZ, WU LY, et al., 2018. Gene expression analysis of bud and leaf color in tea[J]. Plant Physiol Biochem, 107: 310-318.

WU HL, CHEN D, LI JX, et al., 2013. De Novo Characterization of leaf transcriptome using 454 sequencing and development of EST-SSR marker in tea[J]. Plant Mol Biol Res, 31: 524-538.

ZHANG LJ, LIU Q, YANG Q, 2015. Advangage and utilization of tea germplasm resources in qiannan prefecture of Guizhou Province[J]. Chin Tea Proc, 1: 38-41 [张丽娟, 柳青, 杨清, 2016. 贵州省黔南州茶树种质资源优势及开发利用[J]. 中国茶叶加工, 1: 38-41.]